Cognitive Computation sans Representation

[Forthcoming in *Philosophy and Computing: Essays in epistemology, philosophy of mind, logic, and ethics*, Powers, T. (ed.), Springer]

Cognitive Computation *sans* Representation

Paul Schweizer
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
paul@inf.ed.ac.uk

Abstract. The Computational Theory of Mind (CTM) holds that cognitive processes are essentially computational, and hence computation provides the scientific key to explaining mentality. The Representational Theory of Mind (RTM) holds that representational *content* is the key feature in distinguishing mental from non-mental systems. I argue that there is a deep incompatibility between these two theoretical frameworks, and that the acceptance of CTM provides strong grounds for rejecting RTM. The focal point of the incompatibility is the fact that representational content is extrinsic to formal procedures as such, and the intended interpretation of syntax makes no difference to the execution of an algorithm. So the unique 'content' postulated by RTM is superfluous to the formal procedures of CTM. And once these procedures are implemented in a physical mechanism, it is exclusively the *causal* properties of the physical mechanism that are responsible for all aspects of the system's behaviour. So once again, postulated content is rendered superfluous. To the extent that semantic content may *appear* to play a role in behaviour, it must be syntactically encoded within the system, and just as in a standard computational artefact, so too with the human mind/brain - it's pure syntax all the way down to the level of physical implementation. Hence 'content' is at most a convenient meta-level gloss, projected from the outside by human theorists, which itself can play no role in cognitive processing.

1. Introduction

The predominant view in Philosophy, as famously articulated by Fodor (1981) is that "there can be no computation without representation". This assertion is motivated by a particular theoretical stance characterized by two fundamental features. One is commitment to the widely embraced Computational Theory of Mind (CTM), according to which computation (of one sort or another) is held to provide the scientific key to explaining mentality. CTM maintains that cognitive processes are essentially computational processes, and hence that intelligence in the natural world arises when a material system implements the appropriate kind of computational formalism. The second is commitment to the traditionally derived Representational Theory of Mind (RTM), which holds that *representational content* is the key feature in distinguishing mental from non-mental systems.

2. Critique of the Semantic Account of Computation in Physical Systems

The combination of RTM and CTM have given rise to the 'received view' as stated above, which is conveniently expressed in terms of the Semantic Account (SA) of computational implementation, wherein computation in physical systems is stipulated to be the processing

of *representations*, and only physical states that are 'representational' can serve as realizations of abstract formal procedures. However, I will argue that SA is infelicitous for a variety of reasons, and constitutes an unwarranted restriction on the global notion of computation in physical systems. The SA is infelicitous because:

(1) It advocates a departure from the Mathematical Theory Computation (MTC), whereas MTC is the canonical source of our overall theoretical grasp of computation as a cogent and well defined notion. Central to MTC is the intuitive idea of an effective or 'mechanical' procedure, which is simply a finite set of instructions for syntactic manipulations that can be followed by a machine, or by a human being who is capable of carrying out only very elementary operations on symbols. A *definitional* constraint is that the machine or the human can follow the rules without knowing what the 'symbols' are supposed to mean.

There are any number of different possible frameworks for filling in the details and providing a particular specification of the basic idea. Turing's (1936) 'automatic computing machines' supply a very intuitive and elegant rendition of the notion of an effective procedure, but there is a well known variety of alternative frameworks, including Church's Lambda Calculus, Gödel's Recursive Function Theory, Lambek's Infinite Abacus Machines, etc. According to the widely accepted Church-Turing thesis, the notion of computability is nonetheless captured in a *mathematically absolute* sense by the notion of TM computability, and every alternative formalization thus far given of the intuitive notion of an effective procedure has been demonstrated to be equivalently powerful, and hence to be coextensive (and 'non-classical' methods such as connectionism, quantum and analogue computation do not transgress this boundary). The underlying commonality in these frameworks is simply that the rules are finitary and can proceed without any additional interpretation or understanding. As Egan (1995, 2010), Piccinini (2006, 2015) and others have aptly observed, representational content plays no role whatever in MTC

(2) MTC is crystal clear and mathematically precise, while the further restrictive notion of 'representation/reference' invoked by SA is both vague and problematic. Hence this is a retrograde step from clarity and generality to narrowness and potential obscurity. Indeed, given the notorious difficulties in providing a satisfactory rendition of 'representation' in objective scientific terms, SA is in the rather ironic position of promulgating a global restriction on the notion of computation in the physical world that is itself unlikely to be successfully naturalized. More on this point to follow.

(3) Our computational artefacts are the paradigmatic instances of physical computation and can yield any number of counterexamples to SA. As a simple case in point, consider a rudimentary Finite State Machine that accepts some regular language *L*. The FSM operates on strings of uninterpreted syntax and determines whether or not the arbitrary concatenations are grammatically correct. This is an exemplary case of computational processing where absolutely no semantics nor representational content is involved. Or consider a Turing machine intended to compute the values of a particular truth function, say inclusive disjunction. The machine itself is a program for manipulating the symbols '0' and '1' on given input tapes, where '0' is intended to denote False and'1' denotes True. As such, it can easily be *reinterpreted* as computing the truth function associated with conjunction instead of disjunction, simply by flipping the intended reference of the manipulated symbols

so that ‘0’ denotes True and ‘1’ denotes False. There is no independent fact of the matter regarding what these syntactic tokens ‘really mean’ − their referential value is entirely dependent upon a conventional scheme of interpretation which is not itself specified or determined by the computational activities of the Turing machine. The formal behaviour of the device is the same in either case, and the rule governed procedure can be executed with no projected interpretation at all. Indeed, this is precisely what happens with our electromechanical artefacts that successfully run formal programming languages in the absence of any external semantics.

(4) The preceding point highlights an essential flaw related to (1) above: computation is essentially pure *syntax* manipulation, and how the syntax is interpreted is an additional feature not intrinsic to computation *per se*, nor to the successful execution of a formal procedure. SA stipulates that this extrinsic feature is essential, even though the discipline of Computer Science makes no such claim. It has been argued that semantics must be taken into account when individuating computations, because, as above, it is not possible to say which truth function is being computed on the basis of syntactic manipulation alone (Shagrir 2001, Sprevak 2010). And while it is certainly true (and a ubiquitous fact in logic and model theory) that formal syntax underdetermines an intended interpretation, I would argue that the computation *itself* is determined by the formal procedure alone, and whatever semantic value we decide to attribute to the formalism should be seen as a separate question belonging to a distinct level of analysis. Hence SA commits the fundamental error of conflating ‘computation’ *simpliciter* with ‘syntax manipulation under an intended interpretation’. More will be said about this in the following section (see also Dewhurst 2016 for an allied critique of SA).

In response to these infelicities, I would contend that SA is not a viable approach to physical computation. And it's salient to emphasize that the primary reason for making this conflation and attempting to tether the notion of computation to some story about representation does not stem from any issues concerning the general theory of computation itself, but rather is driven by a particular stance within a specialized explanatory project in the *philosophy of mind*. And this is an unduly parochial motivation for promulgating restrictions on the notion of computation in general. All that strictly follows from the conjunction of RTM and CTM is that computation in *cognitive* systems must involve representational content. Hence the more modest received view should instead be the qualified claim that there can be no *cognitive computation* without representation. So RTM plus CTM simply yields the Computational-Representational Theory of Mind (CRTM), which in itself entails nothing about computation in non-mental systems.

Having narrowed the received view from an unwarranted claim about physically implemented computations in general, to its more fitting status as a claim about computations within the specialized realm of cognitive systems, I will now argue that even within this restricted field of application it should not be accepted. There is a fundamental incompatibility between CTM and RTM, which makes CRTM an unstable foundation for a scientific study of the mind. CRTM attempts to wed an ill-defined and pre-scientific criterion of mentality to a formal, mathematical paradigm, whilst I will argue that the two components

are actually quite unsuited bedfellows, and that a serious commitment to CTM provides strong grounds for rejecting RTM.

3. Semantics, Syntax and Formalization

CTM and CRTM have arisen within a pre-existing background context supplied by the rapid development and success of formal methods that began in logic and the foundations of mathematics, particularly in the late 19th century. Traditionally, when we formalize a particular domain of investigation, as in branches of logic and mathematics, we start with our understanding of that domain – with our conceptual grasp of the *intended model*. In such cases the semantic content comes first, and we then devise syntactic systems to capture or reflect crucial aspects of the intended interpretation. The historical roots of this approach can be found in Euclid's project of axiomatizing intuitive geometrical concepts and then rigorously deducing the consequences as theorems.

And in the realm of basic logic we start, for example, with our conceptual grasp of the material conditional as a binary truth function, and then formalize this semantic 'content' with the syntactic derivation rule of modus ponens:

$$\mathbf{A},\ \mathbf{A} \rightarrow \mathbf{B} \vdash \mathbf{B}$$

And contra SA, given this rule it's clearly possible to manipulate symbols in a formal derivation without knowing the truth-table for '→'. Or when formalizing elementary number theory, we begin with our intuitive grasp of the numerical operations of addition and multiplication. We then capture this meaning computationally with the recursive axioms of Peano arithmetic:

$\forall x(x + \mathbf{o} = x)$

$\forall x \forall y(x + y') = (x + y)'$

and

$\forall x(x \cdot \mathbf{o} = \mathbf{o})$

$\forall x \forall y(x \cdot y') = (x \cdot y) + x$

This gives us a mechanical, purely syntactical handle on the intended semantical domain. Ultimately, we can then extend and perfect Euclid's original method to produce formal systems that can be manipulated *automatically*, and hence carry out purely rule governed transformations that *preserve truth* with respect to this domain. Indeed, this is why our computational artifacts are so invaluable: they perform high speed transformations mechanically and automatically that we can then *interpret* with respect to our intended model and thereby discover new truths about that domain. Hence this facilitates the acquisition of vast quantities of new knowledge. In this case we don't say that the realm of elementary number theory itself is computational or formal, but rather that we have provided a *formalization of* our intended model.

So when it comes to Cognitive Science and AI, one *possibility* with respect to computational methods is that (I) we try to formalize the human mind, starting with the assumption that it is some autonomous semantical/representational domain, perhaps comparable to number theory or Euclidean geometry, and we want to devise an automatic formal system reflecting this

domain. Approach (I) is in harmony with the traditional conception of mind that underlies and motivates RTM. Of course, (I) entails nothing about physicalism (and is entirely compatible with dualism), nor does it assert anything about the computational basis of human intelligence. All it claims is that we can (partially) 'capture' a given domain using formal methods, in the sense that rule governed transformations will preserve truth in that domain.

This approach is extremely powerful and general, and has been deployed successfully in a number of areas. For example by adding appropriate non-logical axioms (including Carnapian 'meaning postulates'), Montague Grammar is able to formalize an impressively large fragment of English within a system of higher-order modal. And by axiomatizing the salient natural laws we can formalize scientific theories such a physics. Additionally, in more specialized fields where the dynamical regularities are known we can computationally model a host of phenomena including earthquakes, economies, climate change, hurricanes, particle collisions, protein folding, molecular properties of materials, etc. And such an approach is compatible with 'weak' versions of AI – if we could formalize the human mind, then regardless of its metaphysical status, we could in principle build a computer that simulates this phenomenon. This might then result in artificial humanoid robots that were internally and metaphysically much different than humans, but were nonetheless able to pass a Total Turing Test.

4. The Computational Paradigm

However the CTM view is far stronger and more substantive than this. It contends that (II) human cognition is *itself* essentially computational and that the brain literally implements formal procedures, and so is directly comparable to a computational artifact. Rather than just formally simulating the mind as in scenario (I), this approach attempts to provide a *naturalistic explanation* of mentality in computational terms. So on this view, computation is not a mere simulational or 'engineering' technique, but rather is held to provide the scientific key to cognition. According to the Computational Paradigm, mental states and properties are to be literally described and understood as internal computational processes mediating the inputs and outputs of intelligent behavior. In this manner, computation holds the key not only to explaining mentality in the natural world, but also to the possibility of *reproducing* it artificially, and hence is central to the project of 'strong' AI.

This robust CTM view has a number of theoretical advantages and attractions that are worth reviewing. To begin with, CTM can utilize the relationship between the program level and its realization in physical hardware to provide an elegant solution to the longstanding mind/body problem: according to the mind/program analogy the mind is to the brain as a program is to the electromechanical hardware of a digital computer. In turn, the mind/program analogy offers a compelling solution to the problem of mental causation. Mental processes are analyzed in terms of 'cognitive software' implemented in the neurophysiological 'wetware' of the brain, and any mental event leading to an action will be realized in terms of a physical brain process. For example, the mental event constituted by my desire to raise my right arm is seen as a computational process implemented by my brain, which in turn results in a neuronal

firing that activates the salient nerves controlling my muscles and causing my right arm to rise. No physical conservation laws are violated, and no dualistic 'pre-established harmonies' are required.

Another key virtue of CTM is that, as above, the formal transformation rules can be followed 'mindlessly' i.e. without any outside interpretation or understanding. Classical computation is a process of mechanistically determined transitions from one configuration to the next, and this is very clearly illustrated by the conditional instructions that define a particular Turing Machine. Each instruction is of the form: if in state ***Q****i* reading symbol ***S****j* then perform action ***A****n* (either print a discrete symbol from the prespecified alphabet, or else move one square to the left or right) and enter state ***Q****m* (where *m=i* is permissible). As required by the notion of an effective procedure, the instructions which determine the sequence of transitions on a given input can be executed without any reference to what the manipulated 'symbols' may or may not denote.

So just as in a standard computer, the abstract operations implemented by the brain can be executed without any accompanying semantics. Hence, in principle at least, the 'intentional homunculus' can be fully discharged, and a properly mechanistic and scientific explanation thereby attained. Mentality in the natural world can be accounted for in terms of physically instantiated procedures that do not require any intentional or mentalistic residue. If our mental abilities are essentially computational/formal, then there is no need to invoke any elusive and mysterious phenomena outside the normal posits of natural science (see Schweizer 2001 for further discussion). This is a profoundly significant theoretical gain, and one of the primary scientific strengths of the computational approach. Just as with a standard computational artifact, formal structure and physical law are all that is required. Such a definitive and powerful solution to the problem of the physical, non-dualistic basis of mentality was not even remotely available prior to the 20^{th} century.

Additionally, CTM is in perfect accord with the principles of methodological solipsism and psychological autonomy. Methodological solipsism holds that the study of cognitive processes should consider those processes in abstraction from the environment in which the subject is placed. It's historical roots go back at least as far as Descartes, where skeptical doubt was fuelled by the fact that one's subjective mental realm is compatible with any number of different external circumstances and causes. Analogously, since formal calculi can be manipulated without any appeal to an interpretation, they are internally 'self-sufficient' and independent of the 'external world' of their intended meaning. So in this regard it is fitting to view them in narrow or solipsistic terms. They are incapable of determining a unique interpretation, and cannot distinguish between any number of alternative models.

This fact can be encapsulated in the observation that the relation between syntax and semantics is fundamentally *one-to-many*; any given formal system will have arbitrarily many different interpretations, just as the same narrow psychological state is compatible with a limitless variety of distal sources of input stimuli. To use the classic example introduced by Putnam (1975), if $Oscar_1$ from Earth is transported to Twin Earth, he'll be in exactly the *same*

psychological state when viewing ‘water’, even though on Earth this state was induced by environmental H_2O while now it is induced by *XYZ*.

The term 'methodological solipsism' has potentially misleading connotations as an actual research strategy in cognitive science, since the internal states of an agent will have been profoundly *conditioned by* interaction with its external environment, and hence environmental factors will play a key role in understanding these internal states. The deeper metaphysical import of the notion lies in the 'principle of psychological autonomy' (Stich 1983), which holds that all the properties relevant to the psychological explanation of a given agent must supervene upon its *current, internal physical properties*. In other words, the mental is fully supervenient upon the brain/central nervous system, and thus the external environment can produce changes in the mental states and properties of a subject only insofar as it produces changes in the physical configuration of their brain. Thus the principle of psychological autonomy identifies the boundaries of the cognitive system with the traditionally conceived boundaries of the organism, also coinciding with the causal locus of behavior, the seat of executive control, and the conceptually basic 'individual unit' stemming from the inherited genotype. This localization is of course contested by the extended mind hypothesis of Clark and Chalmers (1998). The current paper will not address the attendant controversy, but will simply advocate the non-extended model, which, along with its many other virtues, also corresponds admirably with the input/output boundaries of a standard computational formalism, and hence is in straightforward accord with CTM.

Thus far CTM exhibits an impressive degree of theoretical integrity and power:

(i) providing an elucidation of the (historically vexed) relation between mind and brain,
(ii) solving the problem of mental causation,
(iii) discharging the intentional homunculus,
(iv) preserving the traditionally conceived input/output boundaries of the organism
(v) providing an account of our cognitive capacities using only the normal resources of the natural sciences
(vi) all within a framework perfectly compatible with the core principle of psychological autonomy.

But now the theoretical waters become seriously muddied…

5. The Postulation of Cognitive ‘Content’

According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. For example, in the scholastic tradition revived by Brentano (1874), the *essential* feature of mental states is their ‘aboutness’ or inherent representational aspect. Searle (1980, 1992) embraces this view with his claim that intrinsic intentionality is the essential attribute possessed by minds, and the feature that must be reproduced in an artefact if (strong) AI is to succeed. And the traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of ‘mental representation’ is adopted as a primary theoretical device. For example, in classical (e.g. Fodorian) cognitive science, Brentano’s legacy is preserved in the view that the properly cognitive level is distinguished precisely by

appeal to representational content. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars, and according to Fodor, it is only when the states of a system are treated as representational, that is, when they are construed as having a content that is *really about something*, that we are dealing with the cognitive level.

Hence the traditionally derived RTM holds that semantic content is essential to mentality. As noted, this view is potentially compatible with approach (I). But Fodor and many others instead attempt to wed RTM with the computational paradigm of approach (II), to yield a theoretical mutation in the form of CRTM. CTM syntax is (multiply) semantically interpretable, and advocates of CRTM would use this opening toehold to try and imbue it with the canonical and venerated 'real content' held to distinguish cognitive from non-cognitive systems. On the CRTM view cognitive agents are described as 'Semantic Engines' – automatic formal systems replete with the unique and privileged interpretation postulated by RTM. Hence the computational syntax of CTM is seen as the 'vehicle' for the essential *content* that is lauded as the hallmark of the mental.

But the first thing to note is that the idea of a 'Semantic Engine' is fundamentally misguided. Only the syntax is mechanized, while the assigned content remains totally inert. As above, the basic purpose of interpreting a formal system is so that we may use the *Syntactic* Engine to discover new truths of our intended model. The model itself does no mechanical work, which is precisely why the *formalization* can supply an epistemic payoff. Since computation is a series of manipulations performed on uninterpreted syntax, the purported content of mental 'representations' is rendered superfluous to the computations that comprise the cognitive processes of CTM. In line with both the principle of methodological solipsism and the definition of an effective procedure, the intended interpretation of internal syntax makes absolutely no difference to the formal mechanics of mind. Thus in the transition to CRTM, one of the prime virtues of a computational approach to mentality has been lost – the discharged homunculus has been smuggled back in. And quite ironically, he now has no real work to do but is just going along for the ride.

5.1 The Narrow Version

Fodor's original notion of mental content was narrow, and this is in line with the principle of psychological autonomy, as well as the orthodox view that content is cognitively and causally relevant to an agent's behaviour, and thus is central to the project of *psychological* explanation. Furthermore, since narrow content is a feature sustained by the individual mental subject, it is also directly in line with the traditional notions of intentionality and mentality from which it derives. Fodor (1994) has since relinquished this view and embraced the notion of wide content. I will argue that both of these positions are mistaken, but will start with a critique of the narrow view, since it is more natural in the context of psychological theorizing, and is more closely aligned with the time-honoured assumptions underlying RTM and consequently CRTM.

According to Fodor's Language of Thought hypothesis (1975, 2008), henceforth LOT, mental processes are explicitly viewed as formal operations on a linguistically structured system of internal symbols. In addition, the LOT incorporates the widely accepted belief-desire framework of psychological explanation, which holds that an agent's rational actions are both *caused* and explained by intentional states such as belief and desire. On the LOT model, these states are sustained via sentences in the head that are formally manipulated by the cognitive processes which lead to actions. Hence propositional attitude states are treated as computational relations to sentences in an internal processing language, and where the LOT sentence serves to represent or encode the propositional content of the intentional state. Symbolic representations are thus posited as the internal structures that carry the information utilized by intelligent systems, and they also comprise the formal elements over which cognitive computations are performed.

Because the tokens of LOT are semantically interpretable and physically realizable in the human brain, they are seen to form a key theoretical bridge between content and causation. So at first pass, this CRTM approach might seem to provide a harmonious theory of the mind/brain, potentially uniting the traditional notion of mental representation with the causally efficacious level of neural machinery. Indeed, this may (possibly?) be why the CRTM approach has such widespread appeal and has become the entrenched orthodox view.
But alas, as argued above, a fatal tension is already built into the foundations of the picture: a central purpose of the symbolic structures is to carry content, and yet, to the extent that they are formal elements of computation, their alleged content is completely gratuitous. Computation is a series of manipulations performed on uninterpreted syntax, and formal structure alone is sufficient for all effective procedures. Indeed, on the foregoing mind/*program* analogy central to CTM, there is a formal procedure or program, while the level of meaning is conspicuously absent. The purported content of mental 'representations' postulated by CRTM is superfluous to the computations that comprise the 'cognitive' processes of cognitive science.

So an obvious move at this point is the one made by Stich (1983) with his Syntactical Theory of Mind (STM) – strip the LOT of its extraneous meaning and let the internal *syntactic* engine churn away on its own. This move is criticized by Crane (1990), who argues that we can't have LOT syntax without attributing semantics. But I think Crane's argument simply reduces to the *epistemological* claim that outside human theorists would not be able to recognize and catalogue the relevant sentences of LOT without first interpreting them. However, even if this were true, it would make no difference to the formal operation of the machinery itself and hence to the actual structure and behaviour of cognitive agents. And what would this attribution of meaning boil down to, other than a case of third person observers assigning some selected sentence of their own public language to a given piece of LOT syntax?

A more serious problem with Stich's STM is that it retains LOT's naive commitment to the common sense categories of belief-desire explanation, and the rather simplistic attribution of privileged and discrete units of innate syntax directly corresponding to our pre-scientific

attributions of mental content. Thus when, in everyday practice, we justifiably ascribe to Jones the belief that lager quenches thirst, both Fodor and Stich would have it that a token of the appropriate mentalese sentence, say 'n%^7 £#~ %&!+', has been duly etched into her 'belief box'. This neuronal implementation of mentalese syntax is then poised to interact with other physically implemented tokens in her desire box to produce assorted forms of rational action, such as standing up and reaching for a pint. Additionally, Fodor would contend that 'n%^7 £#~ %&!+' encodes the very same propositional content as the English sentence 'lager quenches thirst'. Stich rightly notes that this purported content adds nothing to the causal efficacy of the internal syntax so will have no influence on what happens. However, he is still committed to a direct correspondence between common sense public language attributions and fundamental cognitive architecture. I do not wish to become entangled in the 'Folk Psychology' debate at the present time, and will not critically assess this move in terms of its scientific plausibility. Instead I will diagnose what I take to be an underlying conflation between two related but quite distinct theoretical endeavours, and argue that there is a very significant difference between a theory of natural language semantics as opposed to a psychological theory regarding the internal states causally responsible for our input/output profiles. This theme will be further developed below.

5.2 Content Goes Wide

Mental processes and natural language semantics clearly have many intimate philosophical connections, and the foregoing one-to-many relation that underlies the 'symbol grounding problem' has striking and well known consequences for the linguistic theory of meaning. If one accepts the principle of psychological autonomy, then it follows that the mind is too weak to determine what its internal components are 'really about', and this extends to the case of expressions in natural language as well. The famed conclusion of Putnam's Twin Earth argument (Putnam 1975) is that "meanings ain't in the head", and this is because narrow psychological states are incapable of determining the reference relation for terms in our public languages. But rather than abandon natural language semantics in light of the problem, the externalist quite rightly abandons the traditional idea that the intentionality of individuals' mental states provides the foundation for linguistic reference.

Putnam's strategy is to directly invoke external circumstances in the characterization of meaning for public languages. The externalist approach exploits direct indexical and ostensive access to the world, thus circumventing the difficulty by relieving mental states of their referential burden. On such an approach, the object of reference can only be specified by indexical appeal to the object itself, and in principle it *cannot* be determined merely from the psychological states of the language user. Direct appeal to the actual environment and linguistic community in which the cognitive agent is situated then plays the principal role in determining the match-up between language and world. Putnam's strategy offers a viable account of linguistic reference *precisely because* it transgresses the boundaries of the mind as assumed by the explanatory project of (classical) cognitive science. The externalist must invoke broad environmental factors, since nothing internal to a cognitive system is capable of uniquely capturing the purported 'content' of its representations and thereby semantically grounding its internal states. And from this it follows that content is not a property of the representation *qua* cognitive structure, and hence it is not the internal structure nor the

cognitive system itself that provides the theoretical basis for meaning. Indeed, outside factors then do the real work, and the purported *semantical* aspect of cognitive processing is (once again) trivialized.

In light of these considerations, many philosophers (including Fodor 1994, who has now apparently changed his stand again in 2015) have abandoned the traditional notion of narrow content in favour of a wide reading. And while this is an apt move when providing an analysis of the semantics of *public languages*, I would argue that at this point the theoretical projects of natural language semantics on the one hand, and the study of cognitive systems on the other, should be kept clearly distinct. The primary thesis of the current paper is that a genuinely computational approach to cognitive systems has no place for semantic content, since semantics is extrinsic to computation *per se*, and the intended interpretation of a given cognitive formalism will make no difference whatever to the computational processes involved. To the extent that semantic value can *appear* to play a role, it must be syntactically encoded within the system. This is a fundamentally 'internalist' constraint, and is in harmony with the principle of psychological autonomy. Similarly, the type of wide content postulated by an externalist approach to natural language semantics will be causally irrelevant to the behavior of a cognitive system. As Clark and Chalmers (1998) rightly observe, this type of externalism is 'explanatory impotent', in the sense that the difference in the meaning of 'water' on Earth and Twin Earth will make no difference to the behavior of Oscar_1 or Oscar_2 because their narrow psychological states are still identical. And I would take this as a basic limitation on content of any sort, be it narrow or wide, 'naturalized' or otherwise – if difference in the purported content is not manifested via a difference in actual cognitive configuration, then it will have no impact on the activity of the system.

In contrast to Clark and Chalmers, I accept the principle of psychological autonomy, and the associated narrow conception of psychological states. Hence I would argue that, within the context of CTM, nothing that is not part of the *internal syntactical* machinery of a cognitive system can play an efficacious role in its computational cognitive processing. Of course, causation in general is a subtle and controversial topic, and some philosophers, including Shagrir (2001) and Rescorla (2014), reject this narrow conception and hold that wide content *can* play a causal role. It is beyond the scope of the present discussion to engage such claims in detail, but as a broad-spectrum response I would argue that if the framework being advocated is a *computational* approach to the mind, and if the description of the effective procedure in question is held fixed, then the latter is by definition sufficient to account for all well-defined input/output data.

So if wide or external factors are said to have a causal influence on cognitive processes, where this influence is not ultimately manifested in terms of internal syntactic structure, then either one of two possibilities must be the case. The first is that mysterious and completely unspecified 'forces' are 'acting at a distance' to influence the system, and hence more than computational resources are required to account for these mysterious forces, in which case CTM proper has been tacitly abandoned. The other possibility is that the computational formalism in question is being covertly re-described to include formerly 'external' factors which now play a procedurally specifiable role. And in this case the effective procedure in question has *not* been held fixed, but has been implicitly expanded so that elements of the overall context that were previously outside the original input/output boundaries of the formalism have now become *internalized*.

Classical CTM is allegedly committed to viewing the human brain on the model of a piece of computational hardware, and in general the explanatory project of CTM derives from

adopting a perspective that originates within the discipline of computer science. And when designing a piece of electromechanical hardware, professional engineers only take into account (the very well understood) processes of internal physical causation for inventing and constructing the machines used to implement specific computational formalisms. A designed computational artefact is still an open physical system, and is susceptible to any number of non-design-intended outside forces. If it turns out that some external factor is exerting a causal effect on an implemented computation, then this effect must be manifested in terms of a change to the *internal physical structure* of the hardware device, which physical change in turn has an impact on the computational activity in question.

For example, if I run my laptop in a location where it is exposed to direct and very intense sunlight, then the received solar radiation may cause internal components to overheat and this may affect the computational activity of my Windows 10 operating system. External solar radiation is exerting a causal impact on the internal computational processing, but this can happen only because there is an *internal physical change* in the form of components overheating. However, external solar radiation is not itself a relevant 'input' at the computational level of description, and computer scientists and electrical engineers are not in need of any notion of 'wide content' to fully explain the phenomena at issue. Accordingly I would argue that if philosophers advocate an appeal to some mysterious form of 'wide' causation as being necessary for the explanation of *cognitive* systems, then they are implicitly transgressing the bounds of a genuinely computational approach to the mind. And without a well developed account of *how* this causal influence takes place, their accompanying story regarding the dynamics of the underlying medium of physical implementation will be very curious indeed.

6. 'Representation' Talk as a Purely Pragmatic Device

There have been a number of positions advanced in negative reaction to 'orthodox' cognitive science that take anti-representationalism as one of their hallmarks, including dynamical systems theory (e.g Van Gelder 1996), behaviour based robotics (e.g. Brooks 1996), sensory-motor affordances and enactivism (e.g. Noë 2004, Hutto & Myin 2013). and various forms of connectionism. A common factor is that these views all advance some version of the slogan 'intelligence without representation'. In order to locate my position on the salient philosophical landscape, it is important to note that it is *not* anti-representational in this sense. Contrary to the foregoing positions, I would not deny that the most plausible forms of cognitive architecture may well incorporate internal structures and stand-ins that many people would be tempted to *call* 'representations', especially at the levels of perception, sensory-motor control and navigation. So I would be happy to accept things like spatial encodings, somatic emulators, cognitive maps and internal mirrorings of relevant aspects of the external environment. But I would argue that the 'representation' label should be construed in a deflationary and purely *operational* sense, and should not be conflated with the more robust traditional conception from which it derives (see Schweizer 2009 for related discussion). To the extent that internal structures can be said to encode, mirror or model external objects and states of affairs, they do so *via* their own causal and/or syntactic properties. And again, to the extent that they influence behaviour or the internal processing of inputs to yield outputs, they do this solely in virtue of their internal causal and/or syntactic attributes. There is no content involved.

So what I deny is *not* that there may be internal mechanisms that reflect or co-vary with external properties and relations in systematic and biologically useful ways. Instead I would

deny that there is anything *more* to this phenomenon than highly sensitive and historically evolved relations of calibration between the internal workings of an organism and its specialized environmental context. This is a naturalistic description of the system at the 'object' level, and in principle is sufficient to account for all physically tangible interactions and to predict future behaviour on the basis of inputs to the system. And at this level 'representational content' plays no role. Human theorists may then analyze the overall history and environmental context of the system, and from an outside meta level choose to project external 'content' onto various internal structures and processes. But this is a purely extrinsic gloss, and there is nothing about these structures, *qua* efficacious elements of internal processing, that is 'about' anything else. From the point of view of the system, these structures are manipulated *directly*, and the notion that they are 'directed towards' something else plays no role in the pathways leading to intelligent behaviour. Content is not an explicit component of the input, nor is it acted upon or transformed via cognitive computations. In Chomsky's (1995) words,

> There is no meaningful question about the 'content' of the internal representations of a person seeing a cube... or about the content of a frog's 'representation of' a fly or of a moving dot in the standard experimental studies of frog vision. No notion like 'content', or 'representation of', figures within the theory, so there are no answers to be given as to their nature.

The meta-level content postulated by some outside human observers has no computational or causal efficacy, and to the extent that semantic value can *appear* to play a role it must be syntactically encoded within the system. Indeed, just as in a standard computational artefact so too with the human mind/brain – it's pure syntax all the way down to the level of physical implementation.

The internal organization of a given biological system may enable it to achieve certain externally specified goals, and the evolutionary history of the system and consideration of various selectional pressures can shed light on the dynamics of how this has transpired over time. At relatively low functional levels such as perception/navigation/nutrition etc. it is perhaps convenient to abbreviate this by saying that certain internal structures 'represent' external objects and properties, because biologically useful and/or systematic correlations will obtain. And the consideration of wide environmental factors may help us understand *why* certain types of randomly evolved correlations constitute a survival advantage, and even enable us to *look for* mechanisms in organisms which would optimize fitness in such a context.

For example, Shagrir (2014) considers the neural integrator in the oculomotor system. The scientific account is that the system produces eye-position codes by computing mathematical integration over eye-velocity encoded inputs, thereby enabling the brain to move the eyes to the right position. Furthermore, researchers knew beforehand that integration was the function that had to be computed in order for this task to be achieved, and this guided their search for the corresponding neural mechanism. In this case there is compelling reason to view the internal brain process as mirroring or calibrating itself with distal factors in order to successfully control eye position. So one could say that the internal mechanism is a 'representation', if all this means is that there is a clear relation of calibration. But the vital point to note is that *content* plays no role in these mechanisms.

So I would view 'representation' talk as nothing more than potentially convenient shorthand for such basic mechanical facts. And rather than constituting an 'intrinsic' or essential feature distinguishing mental from non-mental systems, the attribution of 'content' is a traditionally derived form of speech that we can sometimes find useful or satisfying. It is a matter of convenience, convention and choice, and does not reveal any fundamental or independent fact of the matter. There isn't a sense in which it's possible to go wrong or be mistaken about what an internal configuration is 'really' about. The pragmatic value of representation talk will depend on different considerations in different applications and contexts of use − so no overarching necessary and sufficient conditions nor univocal meaning. In short, there is no deep issue requiring abstruse and protracted philosophical 'solution'.

7. Behavior versus Meaning

It's clearly the case that many philosophers are of a contrary opinion on the matter, and there have been any number of attempts to 'naturalize' representational content, where the goal is to isolate some unique, privileged and objectively warranted semantic value. The associated literature is vast, and it is well beyond the scope of the current discussion to engage the various positions. As above, my overall view is that within the context of a computational approach to the mind, semantic value *per se* will play no role, whether or not it can be naturalized. However, below I will address a theme introduced earlier in the discussion concerning LOT and the belief-desire framework, and offer a few high level critical considerations regarding the philosophical quest to naturalize the content of propositional attitude states.

The main foundations appealed to in the general naturalization project are causal, informational, and functional/teleological, and one of the predominant current views is the teleosemantic approach stemming from the work of Dretske (1981, 1995) and Millikan (1984, 1986). The technical notion of 'information' does not offer a sufficient basis, given that the mathematically clear and precise analysis provided by Shannon (1948) is purely quantitative and has nothing to say about semantic interpretation (as with MTC!). Hence the additional resources of biological function are invoked, wherein a state of a system is said to represent a piece of information only if this is its proper biological function. So the burden of the intentional homunculus is now shifted to the 'purposiveness' of biological 'design'.

In the case of low level phenomena such as sensation one can play this type of 'naturalistic content attribution game' if one chooses – there is certainly nothing to prevent it. However in the case of high level mental states such as belief and desire the situation becomes much different. A frog may possess an internal structure that tracks the motion of a fly and enables it to snap at the appropriate moment to capture its food, and which has conferred upon it's ancestors a selectional advantage at some stage in the past. In this case there are determinate objects in the external world that can be correlated with the relevant internal mechanisms of the frog. So there is a causally and empirically specifiable process that can be used as the foundation for 'representation' talk, if one is inclined to describe things in such a way.

But in the case of high level mental states such as propositionally individuated beliefs, there is no such determinate object or entity with which an internal processing structure can be mapped or correlated. There is no causal history of mechanical interaction that can be used as a foundation for the story. According to LOT, there *is* an internal configuration, a sentence of

mentalese, which directly corresponds to the belief. But unlike the case of a fly, there is nothing to which this internal sentence can be correlated. Beliefs are typically said to have propositional content, and such content involves a leap of abstraction *many* orders of magnitude beyond sensation or singular reference. There are various attempts to bridge this gap using the resources of control theory, cognitive maps and anticipatory mechanisms. For example, Milkowski (2015) argues that anticipatory mechanisms used to explain the navigational capabilities of rats can provide the basis for the satisfaction conditions required for rich content. And while it may be true that such mechanisms can enable us to explain and predict rat *behaviour*, this still falls far short of providing a foundation for *propositional content*. Rat behaviour must have a purely naturalistic cause, and is itself an empirical phenomenon that we can observe and predict.

In sharp contrast, the propositional content traditionally associated with human belief states cannot be captured by mere behaviour and aspects of the physical system and its environment. The propositional content is expressed via *sentences* in some public language, and the standard approach is to then look for internal structures or processes and relations between agent and external context which can 'bear' or somehow 'naturalize' this attributed content. But alas, there is absolutely nothing in my brain or its interactions with the world that can be identified or correlated with the propositional content of the English sentence, e.g. 'There is no greatest prime number', or 'The ancient Greeks believed in Zeus', or 'It is not possible for Mary to be taller than herself'. There must indeed be some property of me as a biological organism embedded in a particular sociolinguistic community which underlies my disposition to assent to such sentences and deny others, but the *propositional content* itself and its satisfaction conditions far outstrip the realm of physical space-time.

So, as above, I would argue that a line must be drawn between two apparently related but nonetheless quite distinct theoretical projects. There is major separation between a theory of natural language semantics and a psychological theory regarding the internal states causally responsible for our input/output profiles. The former is a highly idealized and normative endeavour, concerned with articulating abstract characterizations which reflect the socially agreed truth-conditions for sentences in a public language. As such, this endeavour has no direct bearing on an essentially descriptive account of the internal mechanisms responsible for processing cognitive inputs and yielding various behavioural outputs, even when we consider the production of *verbal* behaviour, or the common sense attribution of various propositional attitude states *using* natural language.

Granted, in everyday practice, we continually employ sentences of public language to ascribe various content bearing mental states. But this is a projection from the 'outside'. The age-old customs of folk psychology are independent of any assumptions about internal symbols, states or structures. Observable behaviour and context are the relevant criteria, and the truth-conditions for such ascriptions are founded on external, macroscopic and operational considerations. As in everyday life, one can use behavioural and environmental factors to adduce that, say, Jones believes that lager quenches thirst, but this practice makes no assumptions about the nature or even existence of an internal representation encoding the propositional content of the belief. The attribution concerns Jones as an unanalyzed unit, a

black box whose actions take place within a particular environmental and linguistic setting. It gives no handle whatever on postulating hidden internal cogs, levers and teleosemantic functions that generate Jones' actions.

Hence it is vital to distinguish between the semantics of a public language such as English and the internal states and processes of English speaking cognitive agents. *A la* Putnam, there is nothing about the internal states of any English speaker that can determine meanings for a public language. And, *a la* the principle of psychological autonomy, there is nothing about the externally determined semantics of a public language that will impact behaviour, unless this is first manifested via a change to internal mechanisms. The actual workings of the human cognitive system *can* be naturalized, because they constitute a proper subsystem of the natural order. In contrast, the propositional content of public languages cannot be fully naturalized. Propositions are theoretical abstractions, highly normative and idealized extrapolations from human practice that transcend the boundaries of the actual. Indeed, a predominant position in formal semantics is to view them as characteristic functions of *sets of possible worlds*. And the formal definition of such functions is woefully underspecified by the brute facts of physical brain structure and natural selection. Mere terrestrial teleology is one thing, but how on earth could biological evolution select a function designed to yield *XYZ* thoughts on another planet?

8. Conclusion

The efficacy of formal procedures implemented in physical configurations of mass/energy is not affected by the purported presence or absence of meaning, and I would argue that the computational paradigm is thematically inconsistent with the search for content or its supposed 'vehicles'. Instead, the concern of computational models of cognition should be with the internal *processing structures* that yield the right kinds of input/output profiles of a system embedded in a particular environmental context, and with how such processing structures are implemented in the system's physical machinery. These are the factors that do the work and in principle are sufficient to explain all of the empirical data, and they do this using the normal theoretical resources of natural science. Indeed, the postulation of content as the *essential* feature distinguishing mental from non-mental systems should be seen as the last remaining vestige of Cartesian dualism, and computational theories of cognition have no need for a semantical 'ghost in the machine'. When it comes to computation and content, only the vehicle is required, not the excess baggage.

Acknowledgments I would like to thank an anonymous IACAP reviewer for constructive comments, as well as Joe Dewhurst and Alistair Isaac for useful discussion.

References

Brentano, F. (1874). *Psychology from an Empirical Standpoint*.

Brooks, R. (1996). Intelligence without representation. In Haugeland, J. (Ed.), *Mind Design II*, Cambridge: MIT Press.

Chomsky, N. (1995). Language and nature. *Mind*, 104, 1-61.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis* 58: 1, 7-19.

Crane, T. (1990). The language of thought: No syntax without semantics. *Mind & Language* 5, no. 3, 187-212.

Dewhurst, J. (2016). Individuation without representation. *British Journal for the Philosophy of Science*, forthcoming.
Dretske, F. (1981). *Knowledge and the Flow of Information*, Cambridge: MIT Press.
Dretske, F. (1995). *Naturalizing the Mind*, Cambridge: MIT Press.
Egan, F. (1995). Computation and content. *The Philosophical Review*, 104, 181-203.
Egan, F. (2010). Computational models: A modest role for content. *Studies in the History and Philosophy of Science* 41, 253-259.
Fodor, J. (1975). *The Language of Thought*, Cambridge: Harvard University Press.
Fodor, J. (1981). The mind-body problem. *Scientific American*, 24.
Fodor, J. (1994). *The Elm and the Expert*, Cambridge: MIT Press.
Fodor, J. (2008). *LOT 2 The Language of Thought Revisited*, Oxford: Oxford University Press.
Fodor, J. & Pylyshyn, Z. (2015). *Minds without meanings: An essay on the content of concepts*.
Hutto, D. & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA: MIT Press.
Millikan, R. (1984). *Language, Thought, and Other Biological Categories*, Cambridge: MIT Press.
Millikan, R. (1986). Thoughts without laws; cognitive science with content. *The Philosophical Review*, 95 (1), 47-80.
Milkowski, M. (2015). Satisfaction conditions in anticipatory mechanisms. *Biology and Philosophy*, 30, 709-728.
Noë, A. (2004). *Action in Perception*, Cambridge: MIT Press.
Piccinini, G. (2006). Computation without representation. *Philosophical Studies*, 137, 205-241.
Piccinini, G. (2015). *Physical Computation*: *A Mechanistic Account*. Oxford: OUP.
Putnam, H. (1975). The meaning of 'meaning'. In Putnam, H., *Mind, Language and Reality*, Cambridge: Cambridge University Press.
Rescorla, M. (2014). The causal relevance of content to computation. *Philosophy and Phenomenological Research*, 88, 173-208.
Schweizer, P. (2001). Realization, reduction and psychological autonomy. *Synthese*, 126, 383-405.
Schweizer, P. (2009). The elimination of meaning in computational theories of mind. In Hieke, A. and H. Leitgeb (eds.), *Reduction Between the Mind and the Brain*, Ontos Verlag: 117-133.
Searle, J. (1980). Minds, brains and programs, *Behavioral and Brain Sciences* 3, 417-424.
Searle, J. (1992). *The Rediscovery of the Mind*. Cambridge: MIT Press.
Shagrir, O. (2001). Content, computation and externalism. *Mind*, 438, 369-400.
Shagrir, O. (2014). The brain as a model of the world, *Proceedings of the 50th Anniversary Convention of the AISB, Symposium on Computing and Philosophy*. http://doc.gold.ac.uk.aisb50. Accessed 15 July 2015.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423, 623–656.
Sprevak, M. (2010). Computation, individuation, and the received view on representations. *Studies in History and Philosophy of Science*, 41, 260-270.
Stich, S. (1983). *From Folk Psychology to Cognitive Science*, Cambridge: MIT Press.
Turing, A. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceeding of the London Mathematical Society*, (series 2), 42, 230-265.
Van Gelder, T. (1996). Dynamics and cognition. In Haugeland, J. (Ed.), *Mind Design II*, Cambridge: MIT Press.